Sample Size and Heterogeneity Effects on the Analysis of Whole Soybean Seed Using Near Infrared Spectroscopy

Seth L. Naeve,* Rob A. Proulx, Brent S. Hulke, and Tracy A. O'Neill

ABSTRACT

Soybean [*Glycine max* (L.) Merr.] breeders and physiologists are commonly interested in evaluating single plants for seed protein and oil concentration. Moreover, breeders tend to prefer nondestructive techniques so that soybean seed may be retained for variety development. Near infrared spectroscopy (NIRS) technology can be an inexpensive and convenient method for nondestructive analysis of seed; however, many instruments require a 300 to 500-g sample size to operate. Recently, Perten Instruments, Inc. developed a 20 g "breeder's" cup for use with their DA 7200 diode array NIRS instrument. This study was initiated to explore the accuracy and repeatability of this very small cup relative to the standard 300 g cup. While no significant protein bias was discovered, a positive 4 g kg^{-1} bias for oil prediction was found. Seven to eight repeated scans of the breeder's cup were required to increase cup precision compared to a single scan using the standard large cup, potentially due to the large amount of protein and oil heterogeneity within examined seed lots. While the precision of the protein and oil estimation of both cup sizes increased with repeated scans, the low precision noted with the breeder's cup virtually mandates multiple scans or destructive sampling through grinding.

Analysis of soybean seed for protein and oil concentration is fundamental to most soybean breeding and physiology research projects. Instrumentation has been developed to allow researchers to estimate protein and oil in soybean by nondestructive methods on ever decreasing quantities of seed. For instance, a Perten DA 7200 diode array near-infrared spectrophotometer is capable of estimating many quality traits of grains for sample sizes as small as 20 g. This allows for rapid determination of seed quality traits from individual soybean plants or from very small seed lots.

Previous studies have noted the importance of careful sampling when analyzing seed protein and oil content on an individual plant basis, as seed protein and oil concentrations are not uniform throughout the soybean plant. Huskey et al. (1990) found higher oil and lower protein concentrations in seeds from the middle third of the soybean plant when compared to seeds from the upper and lower thirds of the plant. Collins and Cartter (1956) found that seeds from the lower half of the plant were about 0.5% higher in oil and 1% lower in protein than seeds from the upper half of the plant. Bennett et al. (2003) also found higher oil and lower protein content in seeds from the bottom of the plant when compared to seeds from the top of the plant. When analyzed by node rather than by region, there is a linear increase in seed protein content from lower to upper nodes of the soybean plant (Bennett et al., 2003; Escalante and Wilcox, 1993a, 1993b), along with a linear decrease in seed oil content from lower to upper nodes (Bennett et al., 2003). Since nodal variation in protein and oil content appears to be a relatively common occurrence, it is important to sample all seeds from an individual plant when determining protein and oil concentration on an individual plant basis (Collins and Cartter, 1956; Escalante and Wilcox, 1993a, 1993b; Huskey et al., 1990).

Sample size also becomes important when sampling seed gathered from many plants, as seed lot heterogeneity for protein and oil has been noted. Krober et al. (1945) determined that protein and oil differences among 30-g samples of a highly uniform soybean lot were of only "slight significance," while Huskey et al. (1990) found that protein and oil concentration among individual seeds of the cv. Forrest had standard deviations of 29.6 g kg^{-1} and 18.4 g kg^{-1}, respectively. Due to the large seed lot heterogeneity found in their study, Huskey et al. (1990) determined that a sample of 3505 seeds would be required for a mean error limit of 1 g kg^{-1} protein at a confidence level of 95%, assuming a destructive analysis.

Perten Instruments has recently designed a 20 g breeder's cup for use in scanning small seed samples with the Perten DA 7200 NIRS. This cup size is large enough to allow analysis of all seeds from a single average-sized soybean plant. This study was designed to assess the accuracy and repeatability of whole soybean seed protein and oil estimation with this cup in comparison to a 300 g "large" cup, standard for the instrument. Our objectives were to (i) determine whether the breeder's cup would provide the same average protein and oil estimates as the large cup, (ii) identify any differences in scan-to-scan error between cup sizes, (iii) determine whether any scan-to-scan error could be overcome by multiple scans

S. Naeve, R. Proulx, and T. O'Neill, Dep. of Agronomy and Plant Genetics, Univ. of Minnesota, 411 Borlaug Hall, 1991 Upper Buford Cir., St. Paul, MN 55108; B. Hulke, USDA-ARS, Northern Crop Science Lab., 1307 N 18th St., Fargo, ND 58105. Research supported in part by the Minnesota Agricultural Experiment Station. Received 3 July 2007. *Corresponding author (naeve002@umn.edu).

Published in Agron. J. 100:231–234 (2008).
doi:10.2134/agronj2007.0230

Abbreviations: NIRS, near infrared spectroscopy.

of the same sample, and (iv) determine whether sample heterogeneity affects large and breeder's cups differentially due to differences in sample size.

MATERIALS AND METHODS

Seed lots of 10 soybean varieties and unreleased lines with a range in protein and oil concentrations, seed size, and pedigrees were used to approximate the range in seed qualities that many breeding and physiology research projects may encounter. The chosen lines consisted of commercial varieties Asgrow 0803, Asgrow 2107, and Pioneer 91M61, Univ. of Minnesota variety 'Lambert', a non-nodulating isoline of 'Chippewa' soybean, and five lines from the Univ. of Minnesota soybean breeding project (M98-133-20012, M98-297066, M98-308044, M98-324017, and M99-333017). Seed size, protein concentration, and oil concentration for the 10 seed lots ranged from 0.0830 to 0.2430 g seed^{-1}, 372 to 521 g kg^{-1}, and 138 to 212 g kg^{-1}, respectively. Near infrared spectroscopy analysis for protein and oil was performed with a Perten DA 7200 Feed Analyzer (Perten Instruments, Stockholm, Sweden). Four sample cups differing in diameter and capacity were used: a "large" cup measuring 14 cm in diameter with 300 g capacity, a "medium" cup (7.5 cm, 90 g), a "small" cup (6 cm, 50 g), and a breeder's cup (6 cm, 20 g). The large and medium cups were constructed of aluminum painted black, while the small cup was a disposable 59 mL polystyrene condiment cup (P200, Solo Cup Co., Highland Park, IL). The breeder's cup was a Teflon prototype developed by Perten for use in scanning small seed samples, as the light-neutral nature of Teflon reflects little light on its surface.

Study 1 was initiated to investigate whether analysis of soybean seed with the breeder's cup approximated the analysis of soybean seed with the large cup. A subsampling design was used, and the reference sample for this study was a 300-g sample of each seed lot. Reference protein and oil values for the 10 seed lots were first obtained by taking the mean of eight sequential repacked scans in the large cup. Each reference sample was then randomly split into 3, 6, or 15 subsamples for scanning in the medium, small, and breeder's cups, respectively. Each subsample was scanned four times with repacking in its respective sample cup, and subsamples were bulked to reconstitute the reference sample between cup size changes. The large and medium cups are rotated by the NIR instrument during analysis, while the small and breeder's cups are kept stationary during analysis. To provide greater scan accuracy with the small cup and the breeder's cup, an average of two spectral scans was registered for samples scanned with those cups, as recommended by Perten Instruments.

A follow-up experiment, Study 2, was conducted to measure the repeatability of large cup and breeder's cup scans both with and without repacking. The reference sample of each seed lot was first scanned 10 times in the large cup without repacking, and each reference sample was then scanned 50 times in the large cup with a repack between each scan. A random 20-g subsample was then taken from each reference sample and scanned 10 times in the breeder's cup without repacking, followed by 100 subsequent scans in the breeder's cup with a repack between each scan. The scanning of samples both with and without repacking allowed us to estimate the relative contributions of sample heterogeneity and machine error to variation in protein and oil values returned by the NIR machine.

Results of the NIR scans of both experiments were subjected to ANOVA using the GLM procedure of SAS v. 9.1 (SAS Institute, 2002). The hierarchal data structure in Study 1 required that subsamples and scans be treated as nested variables. Data from both experiments were also subjected to a subsampling bootstrap simulation in R-web (Banfield, 2007) to determine the distribution and rate-of-decay of error from the NIR and sampling (Politis and Romano, 1994). Variation due to the effect of soybean seed lot was removed by subtracting the mean seed lot oil or protein content from the original observations. The resulting residuals were then subsampled at n = 50, 20, 10, 5, 4, 3, 2, and 1. For each sample size, the residuals were averaged and recorded. The process of subsampling and computing averages was iterated 10,000 times. The means were plotted on a histogram to determine the distribution of the residuals, and standard deviations of the means at each sample size were computed. Histograms for these data showed a normal distribution, and standard errors of the means decayed at a rate of $1/\sqrt{(n)}$, indicating that the formula $1.96s/\sqrt{n}$ would reliably predict the 95% confidence limits for the mean for any number of scans (n) given the standard deviation estimate (s) from the raw data. Variance and associated standard deviations were also calculated from the raw data within and among seed lots for each cup size, repacking protocol, and trait combination in the second experiment. By dividing the variance of the nonrepacked samples from those that were repacked and subtracting from one, the percentage of the variance that was due to the rearrangement of seeds within the cup was estimated, and expressed as the percentage of variation due to repacking error. Regression analysis was performed with Systat for Windows version 10.2 (Wilkinson, 2002).

RESULTS

From the ANOVA for Study 1 (Table 1), we determined that cup size does not affect mean protein value (P = 0.12), but mean oil values were affected by cup size (P < 0.0001) without a corresponding seed lot × cup size interaction. Table 2 shows mean protein and oil values by cup size across all seed lots. The breeder's cup provided estimates for protein that were nearly identical to that of the large cup, but it allowed overestimation of oil by about 4 g kg^{-1}. The ranges in biases by seed lot were –2.9 to 6.8 g kg^{-1} for protein, and 2.6 to 6.7 g kg^{-1} for oil (data not shown). Overall, the effect of seed lot far outweighed other effects, as 97% of the total sum of squares for both protein and oil was apportioned to seed lot.

The oil ANOVA for Study 1 (Table 1) indicates a significant sample within cup (sample [cup]) effect, and a two-way seed lot × sample [cup] interaction. The two-way interaction was present in the protein analysis as well. This is a result of randomly dividing the reference sample into smaller subsamples for analysis in each of the smaller cups. A significant sample [cup] effect suggests that variation due to seed lot het-

Table 1. Analysis of variance results for protein and oil of 10 soybean seed lots and four near infrared spectroscopy (NIRS) sample cup sizes.

		Protein			Oil		
Source	df	MS	*F* value	*P* > *F*	MS	*F* value	*P* > *F*
		g kg^{-1}					
Seed lot	9	1223.71	1728.18	<0.0001	396.3	1457.91	<0.0001
Cup size	3	1.40	1.96	0.119	5.91	21.76	<0.0001
Sample (cup size)	20	0.97	1.36	0.132	0.51	1.89	0.011
Scans (cup size × sample)	76	0.45	1.05	0.373	0.26	0.94	0.62
Seed lot × cup size	27	0.65	0.90	0.608	0.32	1.17	0.25
Seed lot × sample (cup size)	180	1.31	1.59	<0.0001	0.34	1.23	0.034

Table 2. Mean estimated protein and oil values by cup size across 10 seed lots.

Cup size	Protein	Oil
	g kg^{-1}	
Large	424.8 a†	191.0 c
Medium	425.2 a	191.9 bc
Small	426.8 a	193.1 b
Breeder's	426.9 a	195.3 a

† Means followed by the same letter within a column are not significantly different at $P < 0.05$.

Table 3. Scan-to-scan standard deviation values for both large and breeder's cups where samples were either nonrepacked or repacked. Also indicated are the number of sequential scans with repacking required to meet either 2 or 5 g kg^{-1} confidence limits.

	Large cup		Breeder's cup		±2 g kg^{-1}		±5 g kg^{-1}	
	Nonrepack	Repack	Nonrepack	Repack	Large cup	Breeder's cup	Large cup	Breeder's cup
	g kg^{-1}				no.			
	Protein standard deviation				Scans required for 95% confidence limits for protein			
Avg.	1.39	3.46	1.10	9.07	12.2	87.9	2.0	14.1
Range	1.15–2.18	2.64–4.14	0.69–1.72	5.98–14.48	7.0–17.1	35.7–209.8	1.1–2.7	5.7–33.6
	Oil standard deviation				Scans required for 95% confidence limits for oil			
Avg.	1.37	2.22	1.02	6.21	5.1	42.1	0.8	6.7
Range	0.92–2.18	1.49–2.64	0.80–1.26	2.07–8.62	2.2–7.0	4.3–74.3	0.41–1.1	0.7–11.9

erogeneity may create a large "sampling error" across all seed lots. Due to the two-way interaction noted here, it is likely that this sampling error is not consistent across seed lots for either protein or oil.

As sample heterogeneity appeared to play a role in decisions about design of sampling and analysis protocols, we set out to examine sample heterogeneity in more depth. In Study 2, we examined sample heterogeneity by comparing standard deviations of repeated scans without repacking the sample vs. repeated scans with repacking each time. We found very large standard deviations for both protein and oil when sequential, repacked scans of the breeder's cup were employed (Table 3). Analogous values for the large cup were less than half of those for the breeder's cup. The average standard deviation for sequential, nonrepacked scans of the large and breeder's cups were 1.4 and 1.1 g kg^{-1} for protein and 1.4 and 1.0 g kg^{-1} for oil, respectively. Therefore, cup size had little effect on scan-to-scan error (nonrepacking).

The total variance due to error of NIRS estimation is equal to the sum of repack variance plus variance due to other factors including electronic noise, measurement drift, and differences in particle sizes (in ground samples) (Mark and Workman, 1986). In this study, repacking the sample was responsible for 98 and 95% of total variance noted within seed lots in the breeder's cup for protein and oil, respectively (data not shown). In the large cup, repacking error was responsible for 83 and 56% of the total variance within seed lots for protein and oil, respectively. Since our bootstrap analysis showed that standard errors of the mean for both the large cup and breeder's cup are normally distributed, we were able to calculate the number of repeated scans necessary to reduce the 95% confidence limits of breeder's cup scans to 2.0 and 5.0 g kg^{-1}. The resulting values are shown in Table 3. While there were large seed lot differences, reaching the same confidence interval required seven and eight times the number of repeated breeder's cup scans than large cup scans for protein and oil, respectively. On average, five sequential, repacked scans of the large cup reduced oil standard error to 2 g kg^{-1} for the large cup; however, 42 rescans of the repacked breeder's cup were required to reach the same standard error. Analogous values for protein were 12 and 88 rescans, respectively.

DISCUSSION

The instrument used in this study uses a rotating platter that scans the entire surface area of the specially designed seed cups (the large and medium cups) that hold 300 and 90 g of seed, respectively. The manufacturer states that the scanned area for these two cups is 108 and 44 cm^2. The scanned area of the two smaller cups (small and breeder's) is much smaller (28 cm^2 each) due to the small surface area of the cups. Additionally, these cups do not rotate on the platter and so represent a static flash of the surface of the cup. The differences in scanned area between the breeder's cup and the large cup appear to affect both the accuracy and precision of protein and oil predictions.

Our analysis identified that cup size did affect estimates of oil concentration, with the breeder's cup overpredicting oil by approximately 4.3 g kg^{-1} when compared to the large cup. It is clear that a mathematical bias should be introduced into the oil calibration used for this study for the breeder's cup, as doing so will improve the oil predictability of the breeder's cup. Whether this bias is universally required or is specific to the seed calibration used here could not be determined from this study.

We discovered a very large standard error for soybean samples that were repacked between scans. We determined that a large proportion of this error was due to the act of repacking, while little was due to simple scan-to-scan error. The repacking error was much larger in the breeder's cup than the large

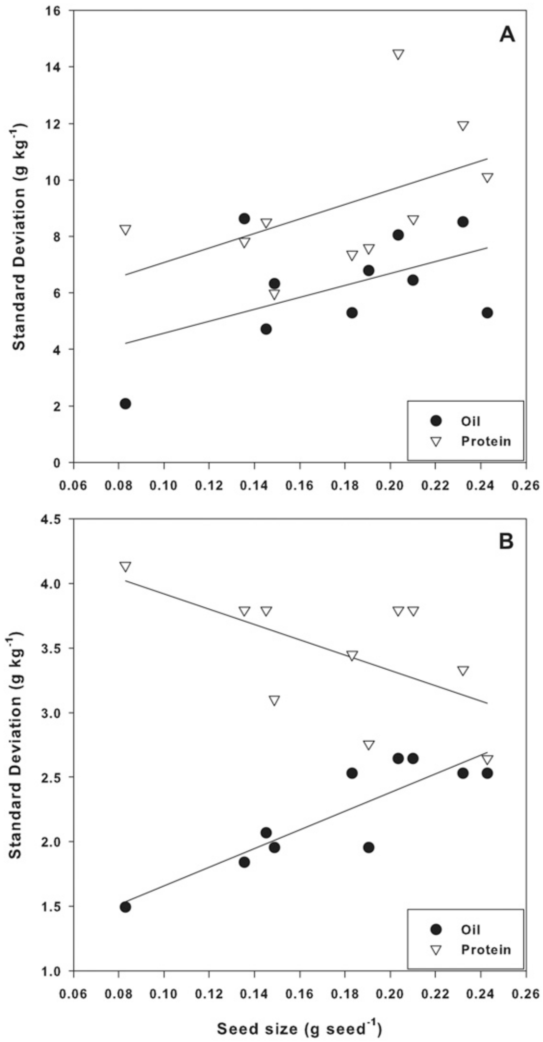

Fig. 1. Effect of seed size (over 10 soybean seed lots) on standard deviation of near infrared spectroscopy (NIRS)-predicted protein and oil concentrations in the breeder's cup (A) and in the large cup (B).

cup. Seed sample heterogeneity undoubtedly played a large role in this effect, based on Huskey et al. (1990) who examined 241 individual soybean seeds from a single seed lot and found ranges in protein and oil concentration from 320 to 520 g kg^{-1} and from 165 to 255 g kg^{-1}, respectively.

In our study, with an average seed size of 0.18 g $seed^{-1}$, 113 and 1 690 seeds were required to fill the breeder's and large cups, respectively. However, the amount of seed scanned within a given cup changes depending on seed size, and seed size differences were correlated with differences in protein and oil standard errors. When the breeder's cup was used, there was a small positive association between seed size and standard deviation of both protein and oil means [r^2 = 0.25 and 0.27, respectively (Fig. 1A)]. As shown in Fig. 1B, the positive association between seed size and standard deviation of oil means was stronger when the large cup was used (r^2 = 0.77), although there was also a negative association between seed size and standard deviation of protein means (r^2 = −0.34). These results indicate that seed size, and therefore number of seeds scanned, did play a role in apparent "sampling error" in this study.

It seems that differences in sample heterogeneity, as described by Huskey et al. (1990) for a single variety, played a role in the large range in standard error difference by seed lot. Seed lots examined here were chosen expressly for genotypic variation. The exact environments where these seed lots were produced are unknown, and it is unknown whether seed lots represented bulking of smaller seed lots across years or locations. Therefore, differences in heterogeneity among seed lots for protein and oil concentration are expected.

CONCLUSIONS

A mathematical bias may need to be included into the calibration equation to account for the significant bias toward overprediction of oil concentration by the breeder's cup relative to the large cup. A large standard error for both protein and oil concentration was discovered for subsequent repacked scans of an individual sample. This error was nearly three times as large for the breeder's cup as the large cup, but can be overcome by averaging sequential scans with repacking each time. The breeder's cup requires seven to eight times the number of sequential scans as the large cup to realize the same confidence limit.

We determined that approximately 100 whole soybean seeds can be evaluated for protein and oil concentration using the breeder's cup and a Perten DA 7200; however, larger standard errors, due primarily to sample heterogeneity, may be evident. This error can be overcome by repeated scans with repacking, resulting in protein and oil measurements equal in precision to measurements made with the large cup. To ensure accurate predictions, the breeder's cup should only be used when limited seed supply necessitates its use, such as for protein and oil analysis of seed from individual soybean plants.

ACKNOWLEDGMENTS

We thank Arthur Killam and Terry Allen for their expert technical assistance, and Jill Miller-Garvin for her critical review.

REFERENCES

Banfield, J. 2007. R-web statistical analysis on the web. Available at http://bayes.math.montana.edu/Rweb/Rweb.general.html (accessed 27 June 2007; verified 30 Nov. 2007).

Bennett, J.O., A.H. Krishnan, W.J. Wiebold, and H.B. Krishnan. 2003. Positional effect on protein and oil content and composition of soybeans. J. Agric. Food Chem. 51:6882–6886.

Collins, F.I., and J.L. Cartter. 1956. Variability in chemical composition of seed from different portions of the soybean plant. Agron. J. 48:216–219.

Escalante, E.E., and J.R. Wilcox. 1993a. Variation in seed protein among nodes of normal- and high-protein soybean genotypes. Crop Sci. 33:1164–1166.

Escalante, E.E., and J.R. Wilcox. 1993b. Variation in seed protein among nodes of determinate and indeterminate soybean near-isolines. Crop Sci. 33:1166–1168.

Huskey, L.L., H.E. Snyder, and E.E. Gbur. 1990. Analysis of single soybean seeds for oil and protein. J. Am. Oil Chem. Soc. 67:686–688.

Krober, O.A., F.I. Collins, and M.J. Demlow. 1945. Sampling soybeans for analysis. J. Am. Oil Chem. Soc. 22:194–196.

Mark, H., and J. Workman. 1986. Effect of repack on calibrations produced for near-infrared reflectance analysis. Anal. Chem. 58:1454–1459.

Politis, D.N., and J.P. Romano. 1994. Large sample confidence regions based on subsamples under minimal assumptions. Ann. Stat. 22:2031–2050.

SAS Institute. 2002. The SAS system for Windows. v. 9.1. SAS Inst., Cary, NC.

Wilkinson, L. 2002. Systat software. v. 10.2. Systat Software, Richmond, CA.